Cell Reports

Supplemental Information

Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells

Iain C. Macaulay, Valentine Svensson, Charlotte Labalette, Lauren Ferreira, Fiona Hamey, Thierry Voet, Sarah A. Teichmann, and Ana Cvejic

Supplemental Table legends

Table S1, Related to Figure 4. Top genes distinguishing each cluster based on random forest feature importance.

Table S2, Related to Figure 6. List of genes with dynamic expression over pseudotime, annotated by which temporal expression group they belong to. Together with enriched functional gene sets for the different temporal expression groups.

Table S3, Related to Figure 7. List of expressed Ohnolog gene pairs annotated by the decision tree classification into Single, Mixed and XOR -Ohnologs.

Supplemental Figure Legends

Figure S1. The gating strategy for sorting cd41-EGFP cells by flow cytometry, Related to Figure 1. First, debris was excluded by forward and side-scatter (A, D). Next, singlets were selected (B, E) and dead, PI positive cells, were excluded (C, F). Finally, autofluorescent cells were excluded from the analysis (G, H). The GFP positive population was split into GFP^{low} and GFP^{high} based on the level of GFP fluorescence (I).

Figure S2. Quality control assessment, Related to Figure 1. Quality control was assessed by analysing the number of detected genes compared to the number of input reads (A) or ERCC content (B). In each plate we sorted 94 cells, leaving two wells per plate without cells. Blue dots represent wells with cells and orange dots show wells without cells. Following sequencing and quality control, 13 cells were removed from further analysis. We excluded data points (cells) with few reads (less than 50,000) and few genes or with high ERCC content. As expected, wells without cells (orange) have ERCC content equivalent to 100%.

Figure S3. Pairwise plots of the four independent components used to represent the data, Related to Figure 2. A) The initial names of the components (“difference_component”, “outlier_component”, “within_large_component”, “within_small_component”) were given based on visual features. The dots, representing cells, are colored white for EGFP^{low} sorted cells and green for EGFP^{high} sorted cells. B) Ward clustering of the cells in ICA space. The clusters (here colored) were used to associate cells to progression along a component where the cluster varies the most.

Figure S4. The gating strategy for sorting cells from clusters 1a/1b/2, 3 and 4 by flow cytometry, Related to Figure 4. A-B) Plots of viable, single cells based on their GFP and PERCP fluorescence from either a non transgenic (A) or Tg(cd41:EGFP) (B) kidney single cell suspension. The GFP^{low} cells (C) can be further split into two groups based on SSC values: GFP^{low}SSC^{high} or GFP^{low}SSC^{low} (D). GFP fluorescence (E) and light scatter (F) properties of each cell coloured based on the cluster it belongs to. G) Stacked column graph showing the proportion of cells from each of the clusters in three different gates named here: GFP^{high}, GFP^{low}SSC^{low} and GFP^{low}SSC^{high}.

Figure S5. May-Grünwald Giemsa staining of cells from clusters 1a/1b/2, 3 and 4, Related to Figure 4. Cd41:EGFP cells were sorted based on GFP and SSC values to GFP^{low}SSC^{high},GFPl^{ow}SSC^{low} and GFP^{high}. Cytospin slides were prepared from sorted cells and stained with May-Grünwald Giemsa. The GFP^{low}SSC^{high} cells are enriched for cells from clusters 1a/1b/2, GFP^{low}SSC^{low} and GFP^{high} cells are enriched for cells from cluster 3 and 4 respectively.

Figure S6. Follow-up experiment, Related to Figure 5. A) Quality control of cells from the follow-up experiment. Out of 288 single cells, 19 were removed from further analysis due to having less than 200,000 sequenced reads, less than 150 detected genes or more than 99.5% ERCC spike-in content in the well. Thresholds were guided by control wells which were either empty or contained 50 cells. B) The data follow a similar pattern as in the original experiment (for comparison please see Figure S3A-B). Pairwise plots of three independent components representing the data from the follow-up experiment. The EarlyEnriched population is confined to the early progression along component 0 (corresponding to within_small_component in Figure S3B) before the switch in component 2 (corresponding to difference_component in Figure S3B). This corresponds to cluster 1a/1b/2 in the original data as expected. GFP^{high} cells from both the kidney and circulation completely overlap, indicating no further differentiation happens after the cells leave the kidney, and vary over component 1 (corresponding to within_large_component in Figure S3B).

Figure S7. The total mRNA content and number of expressed genes per cell are correlated with its differentiation state, not technical properties of the cells, Related to Figure 5. Light scatter properties FSC and SSC, total mRNA content, number of reads and the number of expressed genes in pseudotime. The dots, representing cells, are coloured based on the cluster the cells belong to.

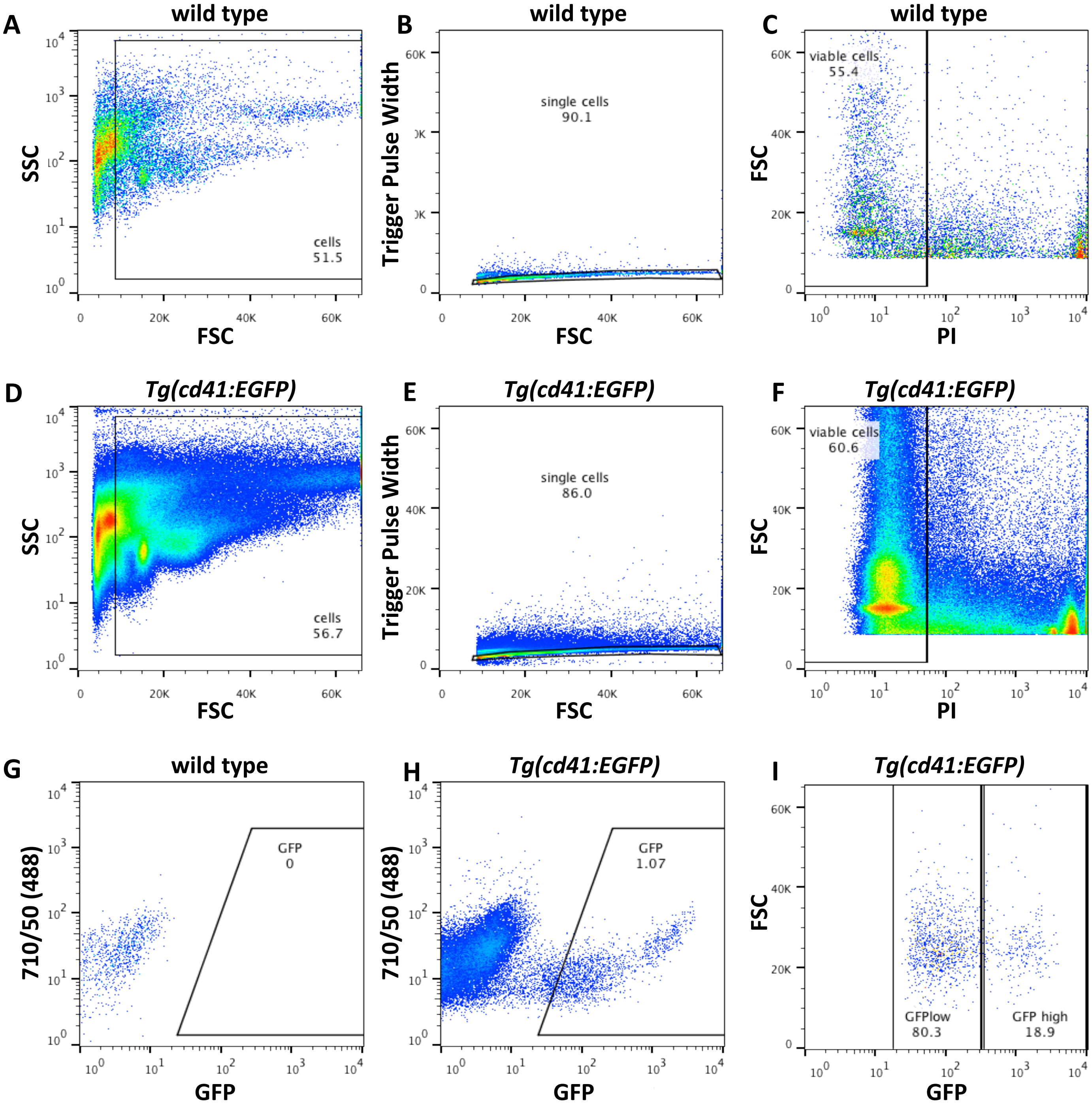
Supplemental Data

Supplemental Data 1. Sample info, Related to Figures 1-7.

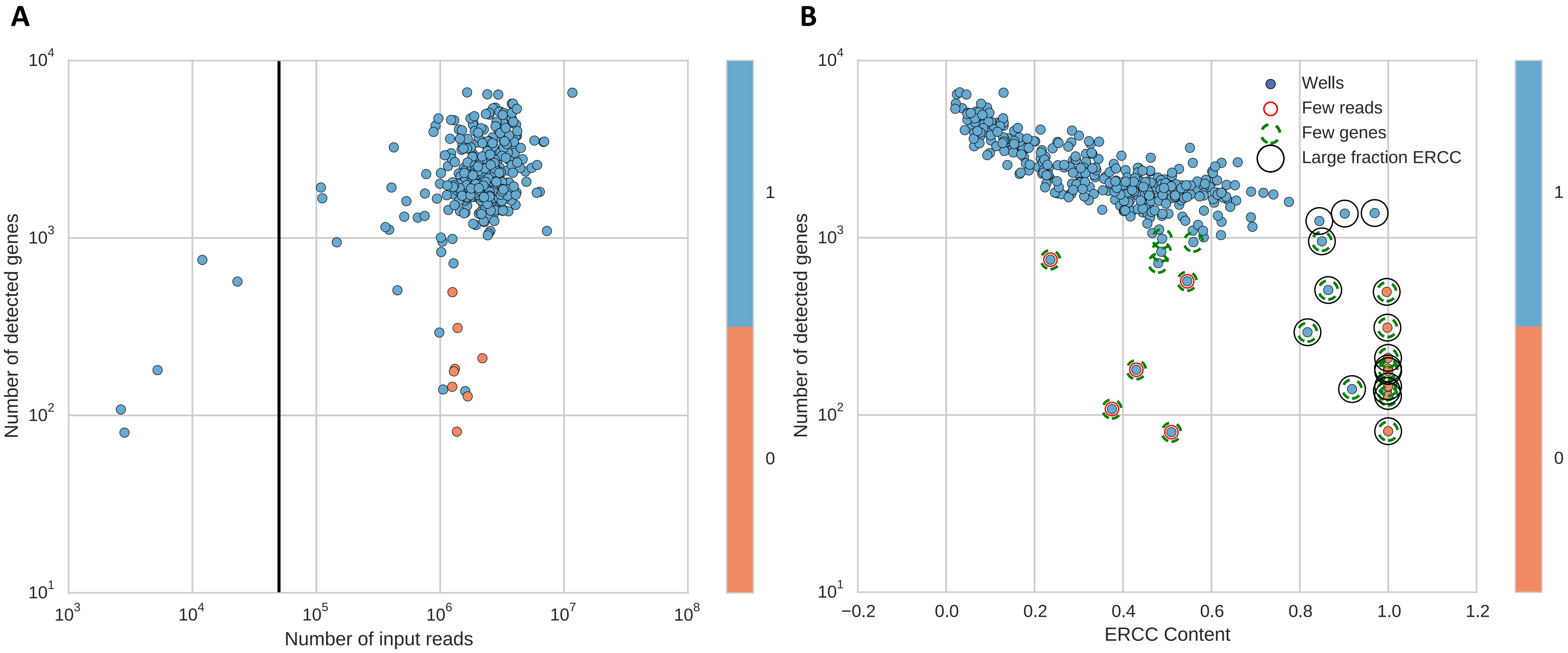
This table contains detailed information about each sample, which was inferred by analysis and used to create most figure panels.

Supplemental Data 2. Analysis files, Related to Figures 1-7.

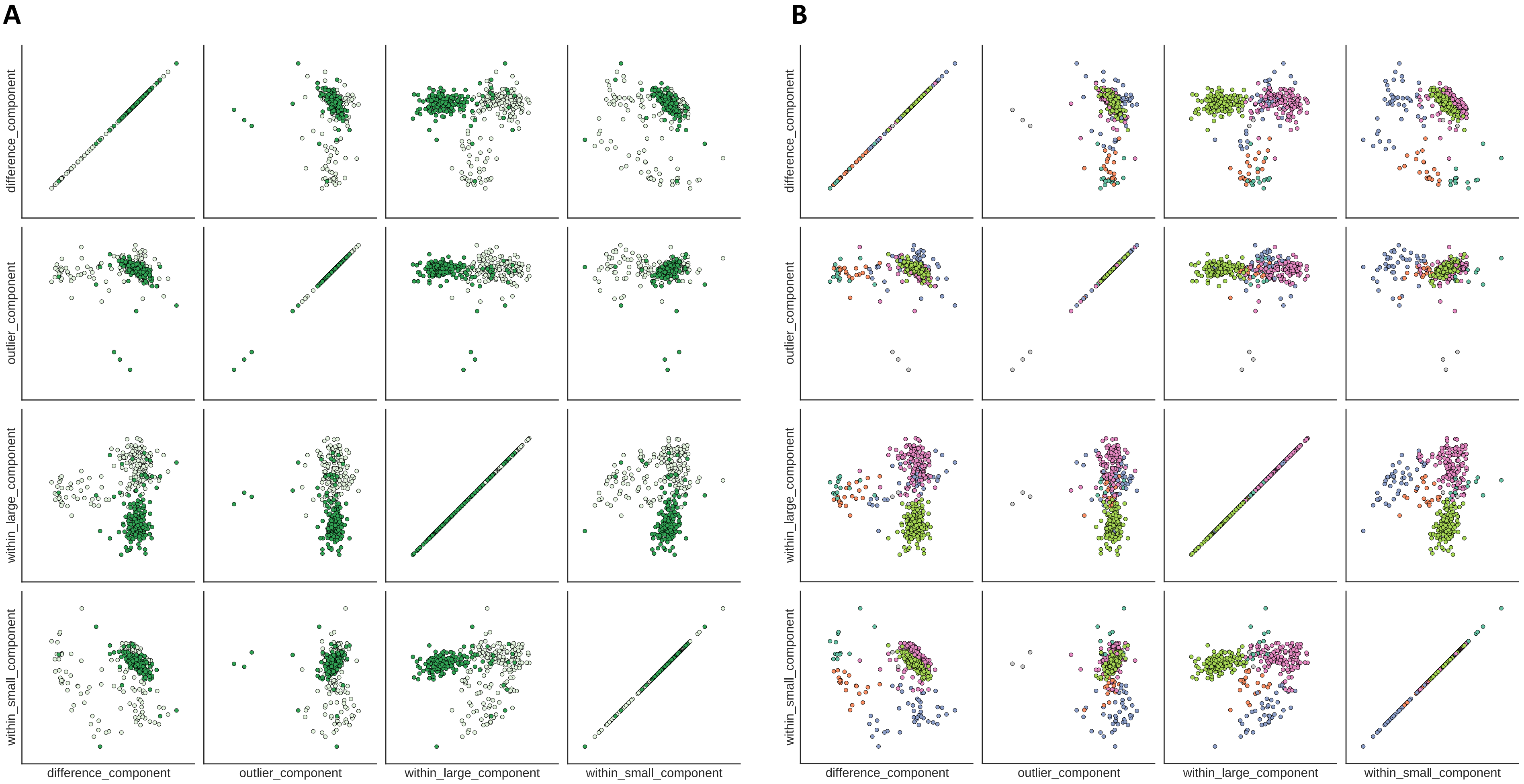
This contains the scripts, in the form of IPython notebooks, to reproduce all analysis and most of the figure panels in the text.


A
wild type
SSC
FSC
cells
51.5
B
wild type
Trigger Pulse Width
FSC
single cells
90.1
C
wild type
FSC
PI
viable cells
55.4
D
Tg(cd41:EGFP)
SSC
FSC
cells
56.7
E
Tg(cd41:EGFP)
Trigger Pulse Width
FSC
single cells
86.0
F
Tg(cd41:EGFP)
FSC
PI
viable cells
60.6
G
wild type
710/50 (488)
GFP
GFP
0
H
Tg(cd41:EGFP)
710/50 (488)
GFP
GFP
1.07
I
Tg(cd41:EGFP)
FSC
GFP
GFPlow
80.3
GFP high
18.9
0
20K
40K
60K
10^0
10^1
10^2
10^3
10^4

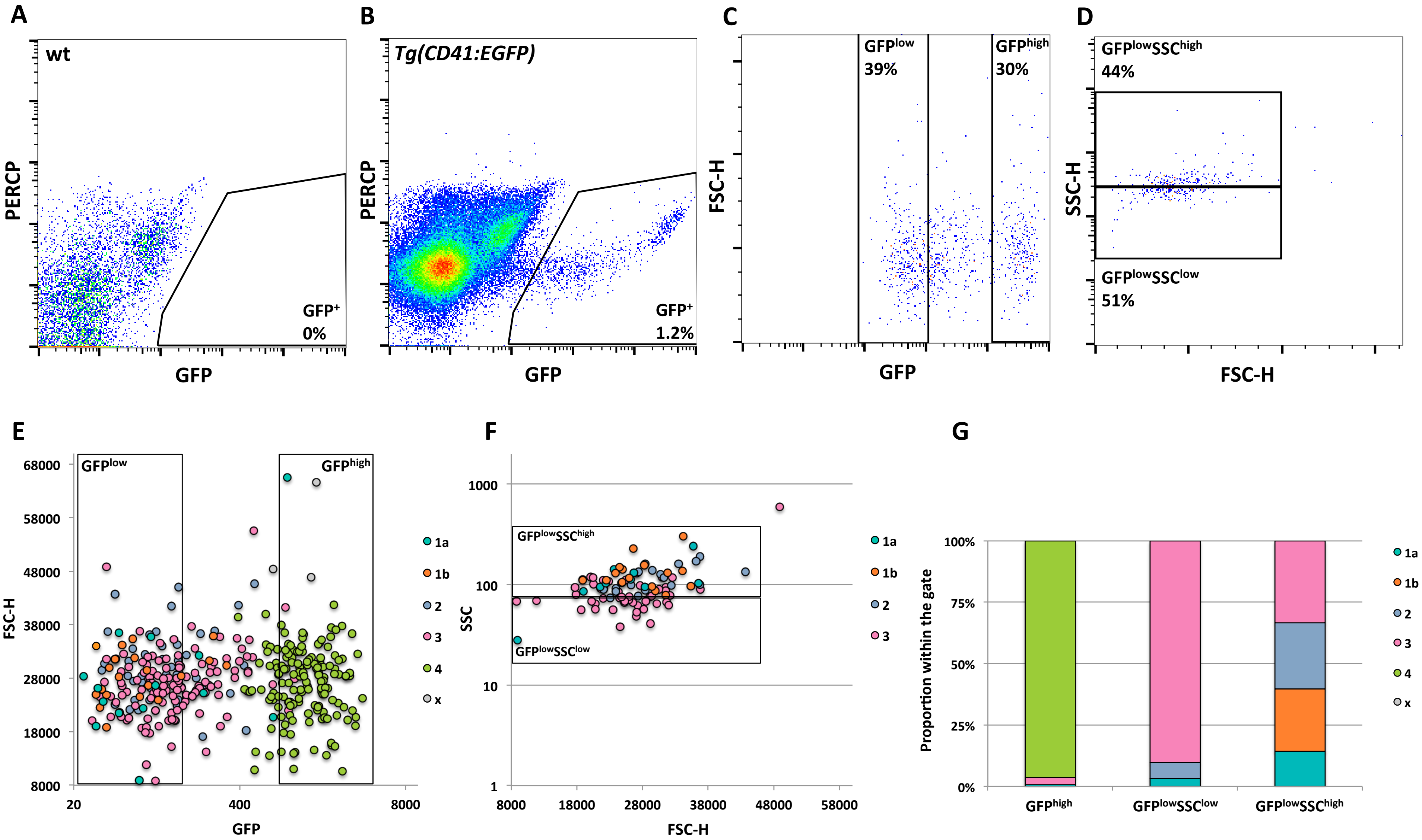
Figure_S1


A
Number of detected genes
10^4
10^3
10^2
10^1
10^3
10^4
10^5
10^6
10^7
10^8
Number of input reads
1
0
B
Number of detected genes
10^4
10^3
10^2
10^1
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
ERCC Content
Wells
Few reads
Few genes
Large fraction ERCC
1
0

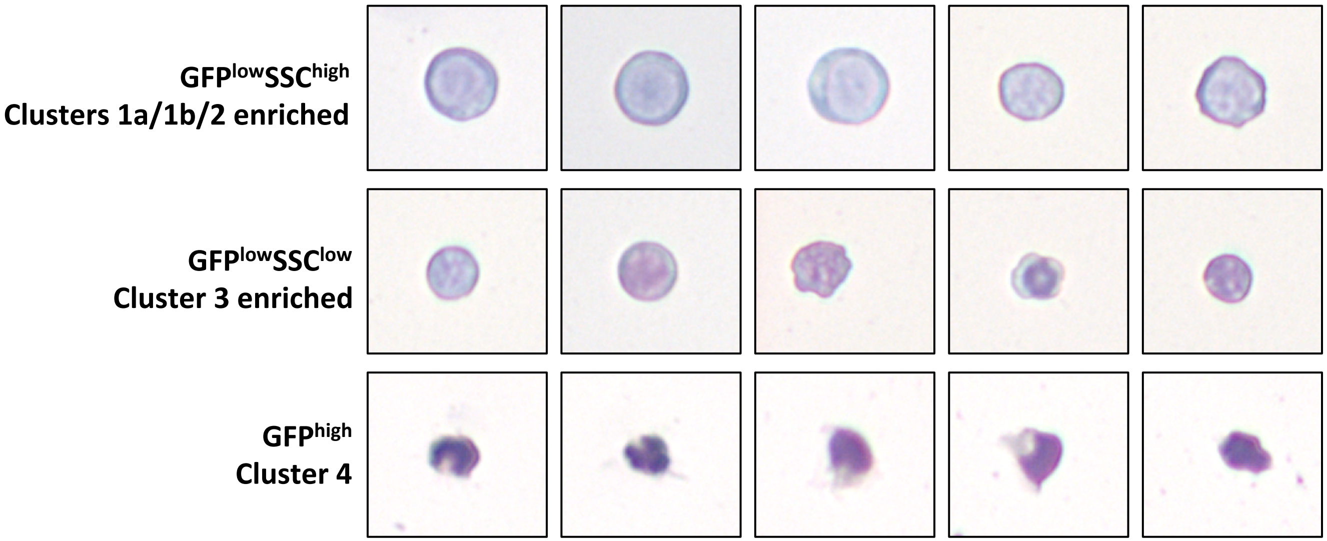
Figure_S2


A
difference_component
outlier_component
within_large_component
within_small_component
difference_component
outlier_component
within_large_component
within_small_component
B
difference_component
outlier_component
within_large_component
within_small_component
difference_component
outlier_component
within_large_component
within_small_component

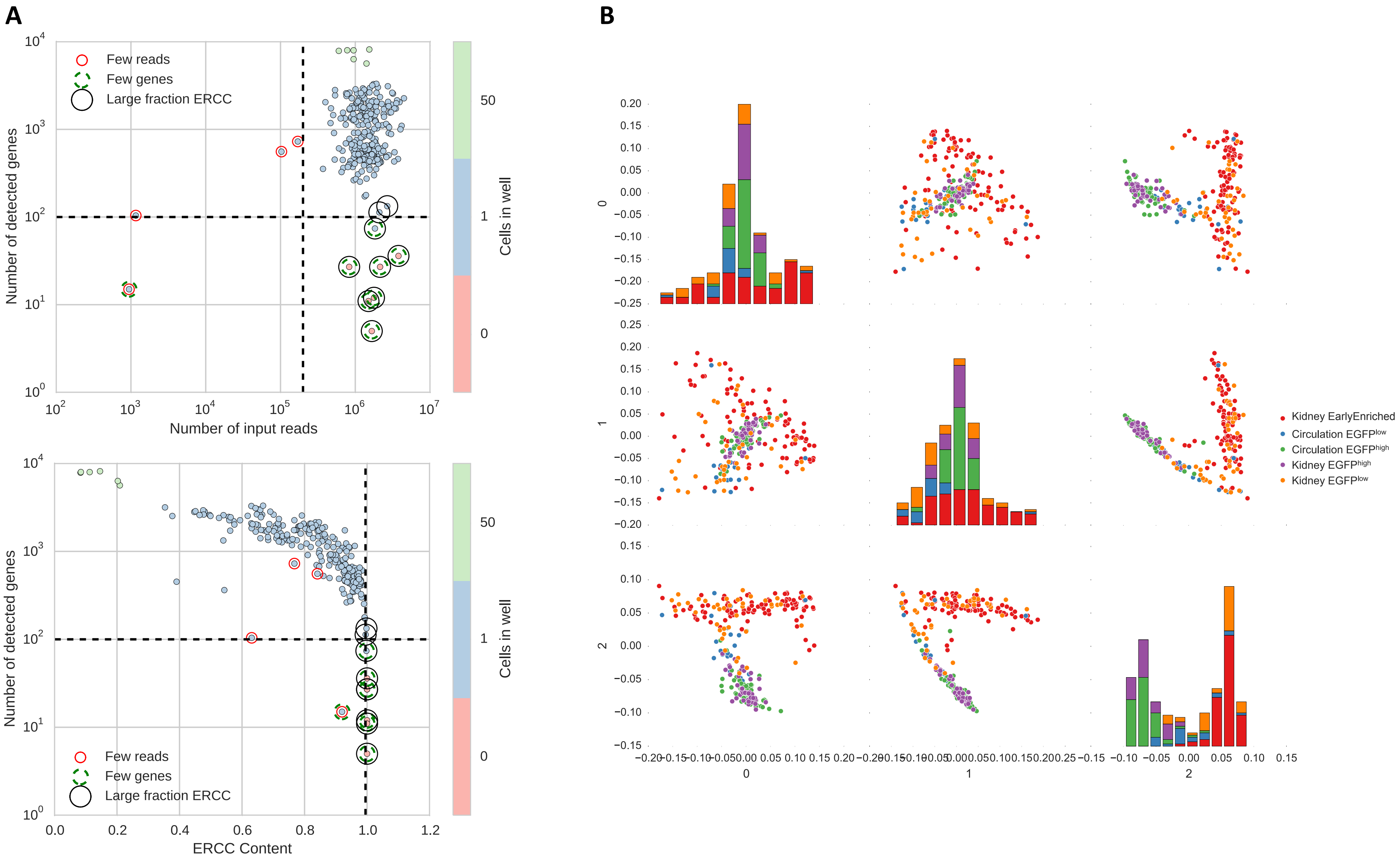
Figure_S3


A
wt
PERCP
GFP
GFP^{+}
0%
B
Tg(CD41:EGFP)
PERCP
GFP
GFP^{+}
1.2%
C
GFP^{low}
39%
GFP^{high}
30%
FSC-H
GFP
D
GFP^{low}SSC^{high}
44%
GFP^{low}SSC^{low}
51%
SSC-H
FSC-H
E
GFP^{low}
GFP^{high}
68000
58000
48000
38000
28000
18000
8000
FSC-H
20
400
8000
GFP
1a
1b
2
3
4
x
F
1000
100
10
1
SSC
GFP^{low}SSC^{high}
GFP^{low}SSC^{low}
8000
18000
28000
38000
48000
58000
FSC-H
1a
1b
2
3
G
100%
75%
50%
25%
0%
Proportion within the gate
GFP^{high}
GFP^{low}SSC^{low}
GFP^{low}SSC^{high}
1a
1b
2
3
4
x

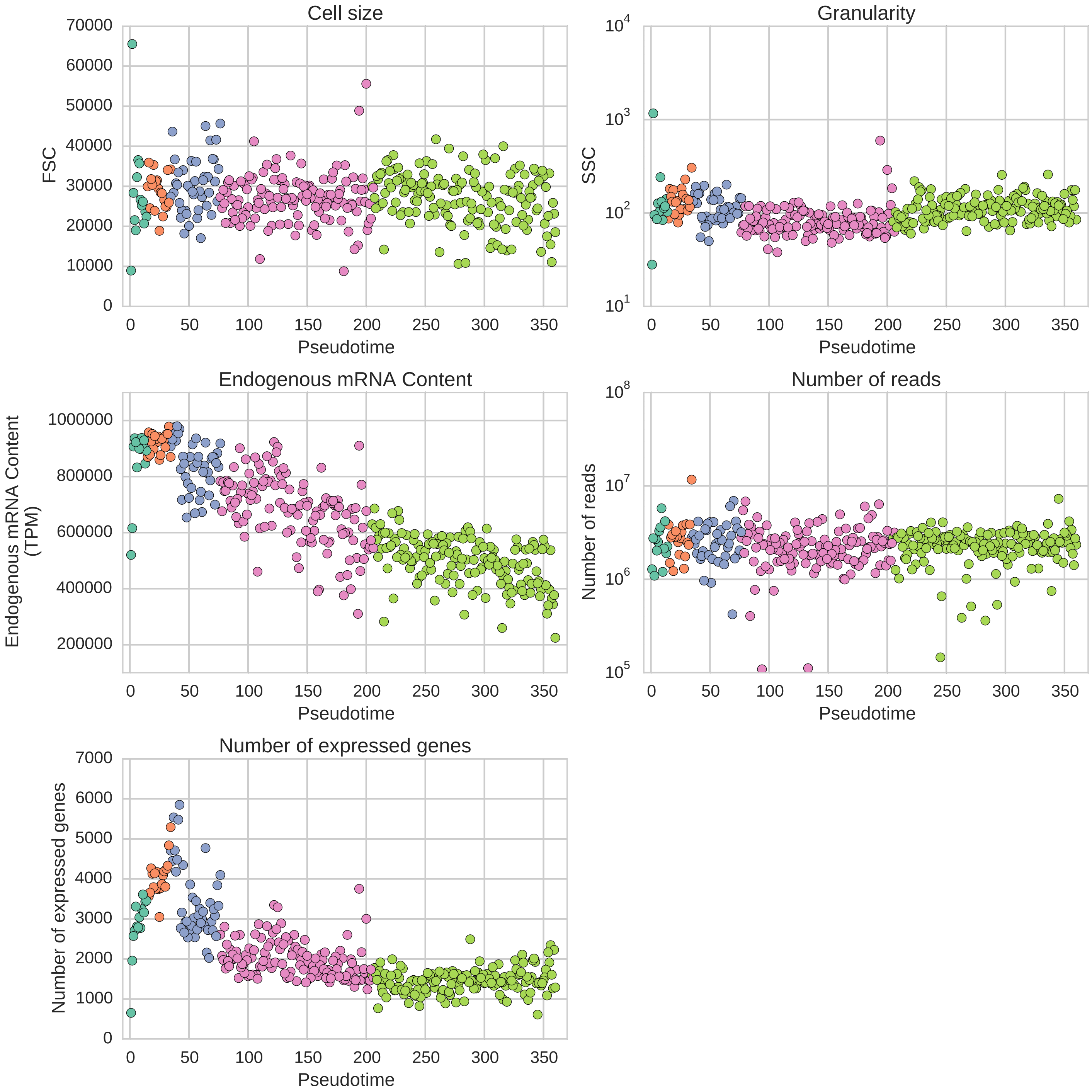
Figure_S4


GFP^{low}SSC^{high}
Clusters 1a/1b/2 enriched
GFP^{low}SSC^{low}
Cluster 3 enriched
GFP^{high}
Cluster 4

Figure_S5


A
Few reads
Few genes
Large fraction ERCC
Number of detected genes
Number of input reads
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
Cells in well
0
1
50
ERCC Content
0.0
0.2
0.4
0.6
0.8
1.0
1.2
B
0
1
2
Kidney EarlyEnriched
Circulation EGFPlow
Circulation EGFPhigh
Kidney EGFPhigh
Kidney EGFPlow
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25

Figure_S6


Cell size
FSC
70000
60000
50000
40000
30000
20000
10000
0
0
50
100
150
200
250
300
350
Pseudotime
Granularity
SSC
10^4
10^3
10^2
10^1
0
50
100
150
200
250
300
350
Pseudotime
Endogenous mRNA Content
Endogenous mRNA Content (TPM)
1000000
800000
600000
400000
200000
0
50
100
150
200
250
300
350
Pseudotime
Number of reads
Number of reads
10^8
10^7
10^6
10^5
0
50
100
150
200
250
300
350
Pseudotime
Number of expressed genes
Number of expressed genes
7000
6000
5000
4000
3000
2000
1000
0
0
50
100
150
200
250
300
350
Pseudotime

Figure_S7